Robust Registration of Astronomy Catalogs with Applications to the Hubble Space Telescope

by

Fan Tian

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering.

Baltimore, Maryland

December, 2018

Abstract

Astrometric calibration of images with a small field of view is often inferior to the internal accuracy of the source detections due to the small number of guide stars in the images. One important experiment with such challenges is the Hubble Space Telescope (HST). A possible solution is to cross-calibrate overlapping fields instead of just relying on standard stars. Following the study in Budavári and Lubow (2012), we use infinitesimal 3D rotations for fine-tuning the calibration but re-formalize the objective to be robust to a large number of false candidates in the initial set of associations. Using Bayesian statistics, we accommodate bad data by explicitly modeling the quality which yields a formalism essentially identical to M-estimation in robust statistics. Our preliminary results on simulated catalogs show great potentials for improving the HST calibration.

Primary Reader and Advisor: Tamás Budavári

Co-Advisor: Amitabh Basu

Acknowledgments

I would like to sincerely thank my advisors Dr. Tamás Budavári and Dr. Amitabh Basu, for their continued support and encouragement throughout this study. They have dedicated countless hours on assisting with my questions and have been more than patient with my mistakes. I could not have imagined having better mentors on my very first journey in research.

I am also grateful to Dr. Steve Lubow and Dr. Rick White from the Space Telescope Science Institute (STScI) for their invaluable discussions and for providing test cases to this study.

Contents

List of Figures

Chapter 1

Motivation

With increasingly available observations from telescopes, astronomy has become one of the most data-intensive fields of study today. The introduction of high-resolution detectors in recent astronomical projects has led to a rapid growth in both data volume and data complexity. To fully utilize information from the vast datasets, it is then essential and often has a great potential for new discoveries to combine observations across multiple wavelengths, at varying time domains, and sometimes between different messengers. Over the last decade, studies in the field of catalog cross-matching have made significant progress using statistical and computational tools. Budavári and Szalay (2008) introduced a reliable framework for symmetric cross-identification of multiple observations based on Bayesian hypothesis testing, which has provided superior results on handling astrometric uncertainties in simulations (Heinis, Bu-

davári, and Szalay 2009). Their methods have also been successfully applied in several studies for cross-matching with unknown proper motions (Kerekes et al. 2010), to incorporate photometry of galaxies (Marquez, Budavári, and Sarro 2014), or studying galaxy clustering (Mallinar, Budavári, and Lemson 2017) and radio morphology (Fan et al. 2013). A review of methods is also available in Budavári and Loredo (2015). More recent studies have also introduced combinatorial optimization methods for cross-identifying associations for 2-way matching (Budavári and Basu 2016) and for N-way matching (Shi, Budavári, and Basu 2017). While the above studies have opened a door for developing new systems and algorithms to address many different problems, new challenges are presented each day to many astronomers for various scientific demands. Among others, a particularly challenging practice arises in cross-matching small images such as those taken by the Hubble Space Telescope (HST).

Unlike large survey projects such as the Sloan Digital Sky Survey (SDSS; York et al. 2000) designed to provide a catalog, HST is not used as a survey telescope in general. For more than twenty years, the HST has been operated under many independent programs targeting specific astronomical objects or sky regions using different detectors. The resultant HST data is a diverse collection of information from all observations made in the past including overlapping exposures at different angles and observations detected in different filters

at different timelines. Cross-matching Hubble images to register the detected sources to a known catalog is more than matching nearby sources as studied in the aforementioned research. It also involves a step of positional adjustment of the images to better align the overlapping sources before matching.

While traditional approach to image registration to the World Coordinate System (WCS; Greisen and Calabretta 2002) standards is promising for large images (Lang et al. 2010), small images such as those taken by the HST are much harder to work with due to the limited reference stars detected in each image. A novel approach taken recently was the project building the Hubble Source Catalog (HSC; Whitmore, Allam, et al. 2016) using the algorithms described in Budavári and Lubow (2012). By rotating images in 3D, they were able to cross-calibrate sources across the HST visits to obtain an improved relative astrometry. With the number of standard stars increased in the aligned images, it also increases the chance to further match these astrometrically corrected images to the larger reference catalogs.

To align the many overlapping HST images, Budavári and Lubow (2012) introduced a 3D infinitesimal rotation vector, which represents the axis and the angle of the rotation for an image. In the context of small corrections, the 3D rotation is also preferred over the traditional transformation performed on the tangent plane, since it avoids many expensive evaluations of the trigonometric functions. The shifts of the images are then determined by minimizing

the separations between paired sources and calibrators that are close on the celestial sphere. This approach essentially arrives at the optimization of a quadratic cost function. The algorithm works effectively when the initial image offset is small, but the issue raises for large residuals that can overpower small values in estimation. The current solution to this problem in HSC is to pre-determine approximately matched pairs using the *pre-offsets* method and a Bayesian likelihood comparison approach (Whitmore, Allam, et al. 2016; Budavári and Lubow 2012). In this study, we propose a new approach that is free from the step of pre-defining nearly matched pairs. To solve for the best transformation, we formulate a robust objective function that can tolerate a large number of erroneous associations in the initial set of candidate matches.

The thesis is proceeded as follows. Chapter 2 presents the detailed robust Bayesian approach and its connection to the M-estimation. Chapter 3 applies the estimation on simulated astronomical catalogs, with discussions on results and limitations of our method. Chapter 4 extends the study to a test case on cross-matching sources from HSC and the *Gaia* Data Release 1 and Data Release 2. Chapter 5 concludes the study with a brief discussion on future plans.

Chapter 2

Methodology

In the following sections, we describe the new method in a simple scenario of cross-calibrating two images. In our procedure, we also follow previous studies and perform estimation on the positional information of astronomical sources extracted from the images. As for simulations, we study on catalogs that contain only point sources representing the detected source directions. In practice, a general approach to work with the HST images is to work on the source directions provided by the source lists in the Hubble Legacy Archive (HLA; Jenkner et al. 2006, Budavári and Lubow 2012). These sources lists are produced by the DAOPhot (Stetson 1987) and the Source Extractor (Bertin and Arnouts 1996) softwares on the combined images within each HST visit (Whitmore, Lindsay, and Stankiewicz 2008). Other than the source directions, the source lists of the HLA also provide information such as the orientations, magnitudes and mor-

phology of the sources detected (Miller, Whitmore, and Jenkner 2008). These additional information can potentially help with verifying and refining the astrometry after correction.

2.1 Pairs and Relative Astrometry

Before carrying out the alignment of sources and calibrators, we need to first determine a set of initial associations between the two catalogs. The initial matchings by pairing all nearby sources within a given search radius R in the three-dimensional space. The distance threshold applied depends on the relative astrometry of the images. While thresholding excludes obvious bad matchings to optimize estimation, a large enough search radius should be used to ensure the inclusion of the maximum number of true associations. We then fix one image and rotate the other to a reference direction relative to the first image. Since the WCS standards are not adapted, we create the set of reference calibrators by using the midpoint directions of the matched pairs. As a result, the astrometic correction of the two images is determined by estimating the rotation of the sources detected from one image to the midpoint directions of the two.

Next we introduce the notations to be used in this chapter. Let the total number of pairs within R to be N_{pairs}. For $q \in \{1, \cdots, N_{\text{pairs}}\}$, we represent the

q-th source direction as $\boldsymbol{r}_q$ with the corresponding calibrator direction as $\boldsymbol{c}_q$, which we then form a set of $(\boldsymbol{r}_q, \boldsymbol{c}_q)$ pairs to be used for calibration. The q-th transformed source in the pair is represented by $\boldsymbol{r}'_q = \boldsymbol{r}_q + \boldsymbol{\omega} \times \boldsymbol{r}_q$ with $\boldsymbol{\omega}$ denoting the 3D infinitesimal rotation vector. With the set of data D on sources and calibrators, we model the cross-matching and registration based on a hierarchical Bayesian framework with parametrization of the factors that jointly describes the transformation.

2.2 Bayesian Formalism

From the set of initial associations, a natural way to think about the contributions from each of the individual pairs to the objective function is that they differ from pair to pair based on their spatial separations. Suppose we have the source-calibrator pairs with small residuals are the "good" members potentially forming the true associations, and for pairs with large residuals are the "bad" pairs, our problem is then to find the registration that maximizes the number of "good" pairs. But the "good" and "bad" pairs are unknown to us, and so is the critical value distinguishes a "good" or a "bad" pair. We thus introduce the binary β variables to represent the two possible states for the set of initial matches, and consider a γ parameter to denote the probability of a pair being "good". We now have the essential parameters to describe both the matching

and the registration and the formal derivation is preceeded as follows.

Let $p(\boldsymbol{\omega}, \beta, \gamma)$ represent the joint prior probability density function (PDF) of the latent parameters. By using the Bayesian inference framework, the posterior probability distribution $p(\boldsymbol{\omega}, \beta, \gamma|D)$ is computed from a prior PDF and a likelihood function derived from data. As our primary interest is the 3D rotation vector $\boldsymbol{\omega}$, we further marginalize the joint probability distribution over other parameters of β and γ. The rotation is then determined as a Bayes estimate to the posterior PDF of

$$p(\boldsymbol{\omega}|D) \propto \int d\gamma \sum_{\beta} p(\boldsymbol{\omega}, \beta, \gamma) p(D|\boldsymbol{\omega}, \beta, \gamma). \tag{2.1}$$

Since the probability for a pair forming a true association can be determined from the number of true associations, the joint prior density in Equation (2.1) has simplified dependencies of

$$p(\boldsymbol{\omega}, \beta, \gamma) = p(\boldsymbol{\omega})\, p(\gamma|\boldsymbol{\omega})\, p(\beta|\boldsymbol{\omega}, \gamma) = p(\boldsymbol{\omega})\, p(\gamma)\, p(\beta|\gamma), \tag{2.2}$$

and the likelihood function is given by $p(D|\boldsymbol{\omega}, \beta, \gamma) = p(D|\boldsymbol{\omega}, \beta) = L(\boldsymbol{\omega}, \beta)$ for

$$L(\boldsymbol{\omega}, \beta) = \left[\prod_{q:\, \beta_q = 1} \ell_q^G(\boldsymbol{\omega}) \right] \left[\prod_{q:\, \beta_q = 0} \ell_q^B(\boldsymbol{\omega}) \right] \tag{2.3}$$

where $\ell_q^G(\boldsymbol{\omega})$ and $\ell_q^B(\boldsymbol{\omega})$ are the "good" and "bad" member likelihood functions

respectively. In practice, a natural choice of the member likelihood function is Gaussian. Here we choose the Fisher distribution (Fisher 1953) - a spherical analogue to the Gaussian distribution - to describe the positional uncertainty of a unit vector direction in 3D. For the observed direction $\boldsymbol{x}$ and the direction of the mode $\boldsymbol{r}$, the PDF of Fisher-distribution is defined as

$$F(\boldsymbol{x}; \boldsymbol{r}, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp(\kappa \, \boldsymbol{r}\boldsymbol{x}) \tag{2.4}$$

with $\kappa = \frac{1}{\sigma^2}$ the compactness parameter for the small astrometric uncertainty σ.

Therefore, a good member likelihood function is then given by

$$\ell_q^G(\boldsymbol{\omega}) = F(\boldsymbol{c}_q; \boldsymbol{r}'_q(\boldsymbol{\omega}), \kappa) \tag{2.5}$$

with transformation $\boldsymbol{r}'_q = \boldsymbol{r}_q + \boldsymbol{\omega} \times \boldsymbol{r}_q$; and a bad member likelihood follows an isotropic distribution of

$$\ell_q^B(\boldsymbol{\omega}) = \frac{1}{4\pi} \tag{2.6}$$

which is the case when $\kappa \to 0$ in Fisher.

Considering the prior probability on β given γ for the good and bad pairs

explicitly such that

$$p(\beta|\gamma) = \left[\prod_{q:\,\beta_q=1} \gamma\right]\left[\prod_{q:\,\beta_q=0} (1-\gamma)\right] \tag{2.7}$$

the joint posterior probability distribution of $\boldsymbol{\omega}$ from Equation (2.1) is then given as

$$p(\boldsymbol{\omega}|D) \propto p(\boldsymbol{\omega}) \int d\gamma\, p(\gamma) \sum_{\beta} \left[\prod_{q:\,\beta_q=1} \gamma \ell_q^G(\boldsymbol{\omega})\right]\left[\prod_{q:\,\beta_q=0} (1-\gamma)\ell_q^B(\boldsymbol{\omega})\right]. \tag{2.8}$$

Marginalizing the joint PDF by summing over the set of $\beta = \{\beta_q\}$ a, Equation (2.8) becomes

$$p(\boldsymbol{\omega}|D) \propto p(\boldsymbol{\omega}) \int d\gamma\; p(\gamma) \prod_{q} \left[\gamma \ell_q^G(\boldsymbol{\omega}) + (1-\gamma)\ell_q^B(\boldsymbol{\omega})\right]. \tag{2.9}$$

To obtain an estimate of an unknown parameter from data, in this case the rotation vector $\boldsymbol{\omega}$, a general approach in Bayesian inference is to compute the mean of the posterior distribution of the parameter (Box and Tiao 1973). And to understand the posterior distribution, an approximate solution is to apply sampling methods such as Markov Chain Monte Carlo (MCMC) (Gamerman et al. 2006). Here instead of finding the Bayes estimate through estimating the posterior probability distribution, we choose an alternative approach of the maximum a posteriori probability (MAP) estimation (Sorenson 1980). That

is, we estimate $\boldsymbol{\omega}$ as the mode of the probability density by maximizing the posterior PDF in Equation (2.9). Moreover, in order to determine the MAP, we also choose to estimate the prior PDF on γ by using the Dirac delta function of $\delta(\gamma - \gamma_*)$. Thus, the integral with respect to γ in the right hand side of Equation (2.9) equals to the likelihood function evaluated at γ_*. As for the choice of γ_*, we follow the discussions in Budavári and Loredo (2015), (also see Budavári and Szalay 2008), and estimate γ_* from the minimum number of sources in two catalogs and the total number of pairs $N_{\rm pairs}$ within the search radius R. Then, for N_1 and N_2 being the number of sources in two catalogs respectively, γ_* is estimated by

$$\gamma_* = \frac{\min(N_1, N_2)}{N_{\rm pairs}}. \tag{2.10}$$

Later we also find that our method is robust to the choice of γ_* which makes the use of Equation (2.10) practical. But one can always refine the estimation after finding the true associations in the corrected catalogs.

With the estimated probability γ_*, and the member likelihood functions, we optimize the posterior distribution through maximizing the joint likelihood function in Equation (2.9) for $\boldsymbol{\omega}$ and obtain the following objective function:

$$\tilde{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \prod_q \left[\frac{\gamma_*}{2\pi\,\sigma^2} \exp\left\{ -\frac{\left[\boldsymbol{c}_q - (\boldsymbol{r}_q + \boldsymbol{\omega} \times \boldsymbol{r}_q)\right]^2}{2\sigma^2} \right\} + \frac{1-\gamma_*}{4\pi} \right]. \tag{2.11}$$

2.3 Connection to *M*-estimation

When all pairs are "good", i.e. $\gamma_* = 1$, Equation (2.11) yields the least squares problem as introduced in Budavári and Lubow (2012). As the fraction of good pairs decreases, the effective likelihood function gains heavier tails making the optimization more difficult. To find the optimum, we borrow ideas from robust statistics (Huber 1981) to reformalize our objective function in Equation (2.11).

Let the separation between the q-th source-calibrator pair to be $\boldsymbol{\Delta}_q = \boldsymbol{c}_q - \boldsymbol{r}_q$. For any given γ_*, instead of maximization, we minimize the negative logarithm of the likelihood function in Equation (2.11) and arrives at the following objective function

$$\tilde{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \sum_q \rho\left(\frac{|\boldsymbol{\Delta}_q - \boldsymbol{\omega} \times \boldsymbol{r}_q|}{\sigma}\right) \quad \text{with} \quad \rho(t) = -\ln\left(\frac{2\gamma_*}{\sigma^2}\, e^{-t^2} + 1 - \gamma_*\right) \tag{2.12}$$

and $\boldsymbol{\Delta}_q = \boldsymbol{c}_q - \boldsymbol{r}_q$. As illustrated in Figure (2.1), this ρ-function is quadratic for small residuals, but constant for large values - limiting the contribution of bad pairs to the objective. We note that ρ is a function of t^2 only and this problem formally is much like the M-estimation in robust statistics (Maronna, Martin, and Yohai 2006). The solution exists requiring the gradient of the objective function equals to zero. Since no closed formed solution exists for an M-estimation, a general approach is to use an Iteratively Reweighted Least

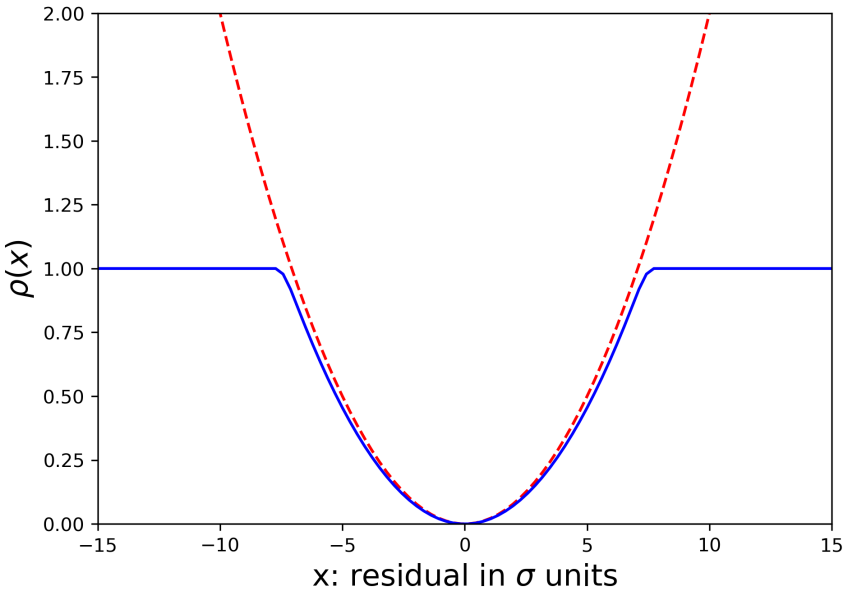

Figure 2.1: The robust ρ-function (*solid blue line*) limits the influence of outliers in comparison to a quadratic objective (*dashed red line*).

Squares (IRLS) method (Maronna, Martin, and Yohai 2006). In this study, we solve the problem by iterating between (1) solving for $\tilde{\boldsymbol{\omega}}$ using $A\,\tilde{\boldsymbol{\omega}} = b$ with

$$A = \sum_q \frac{w_q}{\sigma^2}\left(I - \boldsymbol{r}_q \otimes \boldsymbol{r}_q\right) \quad \text{and} \quad b = \sum_q \frac{w_q}{\sigma^2}\left(\boldsymbol{r}_q \times \boldsymbol{c}_q\right)$$

assuming constant w_q weights, and (2) re-evaluating those weights based on $\tilde{\boldsymbol{\omega}}$ as

$$w_q = W\left(\frac{|\boldsymbol{\Delta}_q - \tilde{\boldsymbol{\omega}} \times \boldsymbol{r}_q|}{\sigma}\right) \tag{2.13}$$

with $W(t) = \rho'(t)/t$. We find this procedure converges quickly in practice.

Chapter 3

Simulations

To understand the performance and the limitations of our new method, before applying to the real observations, we first investigate simulations where the ground truth is known. In the test setting, the mock objects and their simulated detections are created in realistic scenarios to the HST observations.

3.1 Mock Universe and Catalogs

Our mock objects generated in a small field of view are point sources with random directions. Each point is taken as an unit vector representing the pointing direction to the actual coordinates on the celestial sphere. In addition to the directional information, each object is assigned with a random stellar property u_{01} drawn from a standard uniform distribution between 0 and 1.

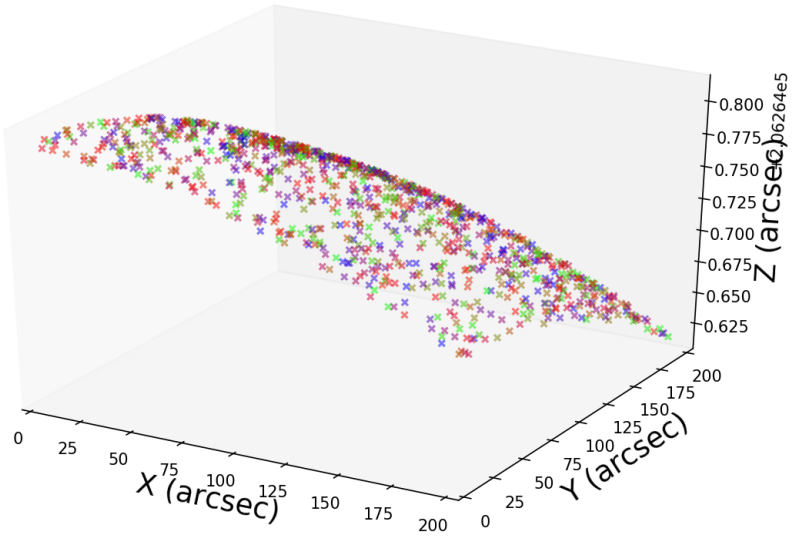

Figure 3.1: Scatter plot of mock objects colored by stellar properties of courses.

Figure (3.1) is a 3D representation of the mock objects colored by their physical properties.

From the mock universe, catalogs are generated in pairs by (1) assigning random perturbations to the mock objects with a chosen astrometric uncertainty; (2) selecting overlapping sources from transformed catalogs by a interval constraint on the source property u_{01}; and (3) transforming one of the catalogs from the pair with a random rotation vector ω drawn from a normal distribution, and the other with a mirror vector of $-\omega$. The estimation is thus performed on the pairs of catalogs aiming to recover the rotation vectors ω used in generating the catalogs. As an example to the simulated catalogs, the left panel of Figure (3.2) shows a 2D projection of two catalogs (point sources colored with blue and orange colors) generated from the mock universe. The

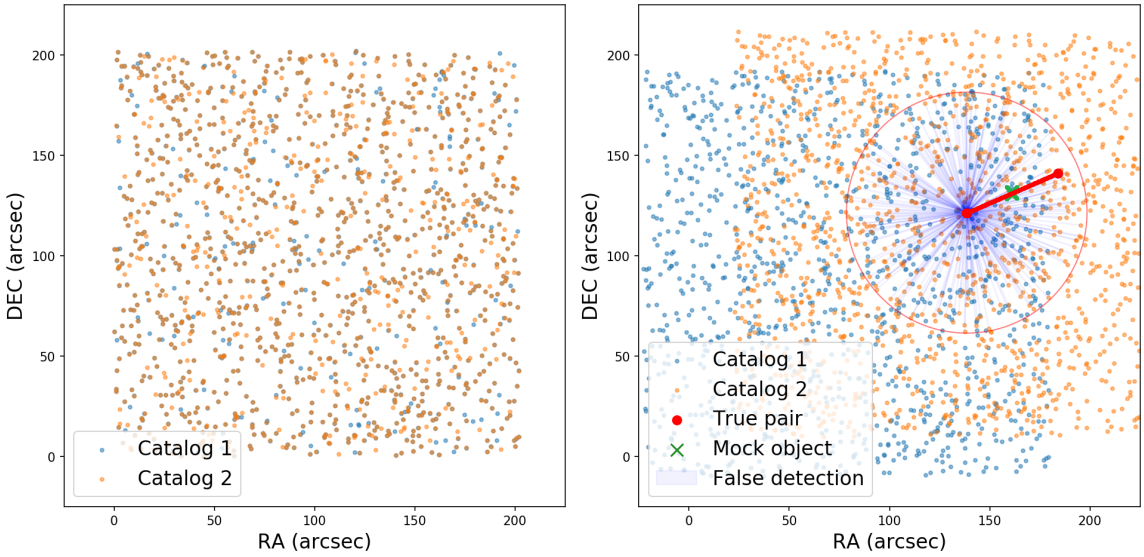

Figure 3.2: 2D projection of catalogs. Left panel: Source directions without transformation. Right panel: Transformed catalogs. Ground truth shows for a single source (red center), there is one true matching pair (red line) with underlying object (green cross), but many false matchings (blue line) within search radius R.

right panel of Figure (3.2) represents the same catalogs with transformation applied. Additionally, Figure (3.2) (right panel) also shows the challenges for matching when the image offset is large a large search radius (red circle) is applied. For the highlighted source (red center) in catalog 1, the ground truth indicates that there is only one true matching in catalog 2 (red line) such that this pair corresponds to the same underlying astronomical object (green cross). The blue lines represent the many false matchings for the singled-out source within the given search radius R. Our estimation is therefore performed to align the true associations to the object's direction while tolerating the large number of bad matchings.

3.2 Numerical Effect

Our simulated catalogs started with a smaller field of view, a small number of sources and an astrometric uncertainty of 0.1 arcsec to effectively experiment with the implementation of our new method. As we progressed from successful testings on small catalogs, the final setting targets the realization of the HST images taken by the Advanced Camera for Surveys in the Wide Field Channel (HST/ACS/WFC; Lucas et al. 2018). The filed of view of ACS/WFC images is 202 arcsec $\times$ 202 arcsec with approximately 1500 source detections in each image. The astrometric uncertainty used adapts the HST positional accu-

racy of $\sigma \sim 0.04$ arcsec. The change made in the simulation setting has allowed us to identify a major issue in our method - we found that changing the astrometric uncertainty parameter σ affects the estimation results significantly. This is later investigated to be a numerical issue in the objective function. For the vary large residuals, a small σ value causes the exponential term in the ρ-function asymptotically approaches to zero. When summing over all pairs, we are in fact summing over a list of zeros that can lead to an early arrival to a local minimum instead of finding a correct global minimum. To elaborate, we represent the objective function for different σ values in Figure (3.3). For small σ values, sum of the ρ-functions has more sharp peaks such that the estimation is more likely to be trapped at a local minimum. As σ increases, the objective function is smoothed out, which hence increases the chance of finding a correct estimation. Therefore, our current approach to address this numerical issue is to artificially assign a large enough value to σ at initial steps of the iterations. After a certain number of iterations, we converge σ to the actual astrometric uncertainty of the catalogs and proceed estimation until convergence.

3.3 Algorithm

Our current implement of the new robust method is proceeded as shown in Algorithm 1. For every two catalogs, our inputs are the matched source-

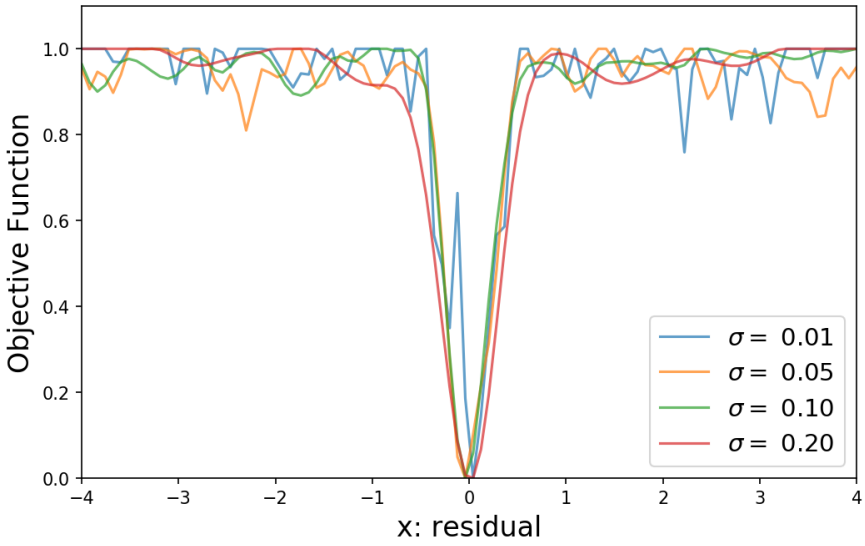

Figure 3.3: Changing σ affects the smoothness of the objective function

calibrator pairs, the search radius and the uncertainty σ. By initializing the weights to be all ones, we compute an initial estimate of $\tilde{\omega}$ and use it to refine weights and repeat. The initial σ adapted is a fraction of the search radius, which we used in this study is $\sigma_0 = \frac{1}{3}R$. For the total number of iterations to be T, we decrease σ_t for a number of iterations of T_* such that $T_* < T$. When σ_t is converged to the right value, we start to evaluate the objective function value with the actual σ and proceed to an automated stop of the iterations. The termination criterion applied in our algorithm is given by the objective function value evaluated at the current and the previous estimates of $\tilde{\omega}$. That is, for $F(\boldsymbol{\omega}) = \sum \rho(\boldsymbol{\omega})$ being the objective function, we stop the procedure when $\frac{|F(\tilde{\omega}_t) - F(\tilde{\omega}_{t-1})|}{|F(\tilde{\omega}_t)|}$ is less than a small threshold of τ_{stop} We used $\tau_{\text{stop}} = 10e^{-14}$ for the double precision of Python being $10e^{-16}$

Algorithm 1 Robust estimation

Input:
Source-calibrator pairs, Search radius R,
Astrometric uncertainty σ
Initialize:
$\{w_q\}_0 \equiv \mathbf{1}$, $q = 1, \ldots, N_{\text{pairs}}$
$\gamma_* \leftarrow \min(N_1, N_2)/N_{\text{pairs}}$
$\sigma_0 \leftarrow \frac{R}{3}$; $\sigma_{T_*} \leftarrow \sigma$
$F(\boldsymbol{\omega}_0) \leftarrow inf$ ▷ Initialize a large objective function value
stopping tolerance $\boldsymbol{\tau}_{\text{stop}}$
for $t = 0$ to T with $T_* < T$ **do**
    Compute $\tilde{\boldsymbol{\omega}}_t$ by solving $A\tilde{\boldsymbol{\omega}}_t = b$
    Update weights by $\{w_q\}_t \leftarrow W(\tilde{\boldsymbol{\omega}}_t)$
    Update $\sigma_t^2 \leftarrow \frac{T_* - t}{T_*}\sigma_0^2 + \frac{t}{T_*}\sigma_{T_*}^2$
    if $t > T_*$ **then**
        if $\frac{|F(\tilde{\boldsymbol{\omega}}_t) - F(\tilde{\boldsymbol{\omega}}_{t-1})|}{|F(\tilde{\boldsymbol{\omega}}_t)|} < \boldsymbol{\tau}_{\text{stop}}$ **then**
            break
        end if
    end if
end for
return $\tilde{\boldsymbol{\omega}}_T$

3.4 Results and Discussion

Applying the same simulation setting but with different rotation vectors, we test the new algorithm on a set of catalogs with different search radius and different image offsets. As a comparison, the least squares method is also tested under the same conditions. The accuracy of the method is reported by comparing the initial image offset and the offset after correction.

We compare both the least squares method and our robust method in two ways. We first compare the estimation accuracy of two methods for images with a small offset. As shown in Figure (3.4) (top panel), for two images with an initial offset of approximately 0.1 arcsec, when increase the search radius, both methods recover the correct rotation for $R < 1$ arcsec. For $R > 1$ arcsec, the least squares method starts to break down. Our new robust estimate, on the other hand, can find the accurate rotation vector under large R. The bottom panel measures γ as the fraction of the number of true pairs to the number of pairs within R. This reinforces the fact that the least squares estimation is less robust to extreme residuals. Furthermore, we have compared two methods for images with large offsets and the results shown in Figure (3.5). To draw a fair comparison, we applied a search radius to be just a few σ above the initial offset. With a small initial offset ($< \sim 0.3$ arcsec), both our robust estimate and the least squares estimate correct the astrometry to approximately σ. As the initial offset increases to above 0.3 arcsec, the least squares algorithm fails

to find a correction. This finding also coincides with the current limitation to the algorithm implemented in HSC. Moreover, our robust estimate is accurate for offsets up to 60 arcsec (1 arcmin). Beyond 1 arcmin, neither method can satisfactorily recover the rotation. Although images with offsets larger than one arcmin are rare cases of the HST, we would still want to address these scenarios. The previous approaches on pre-determining likely associations are still preferred under the very large offsets.

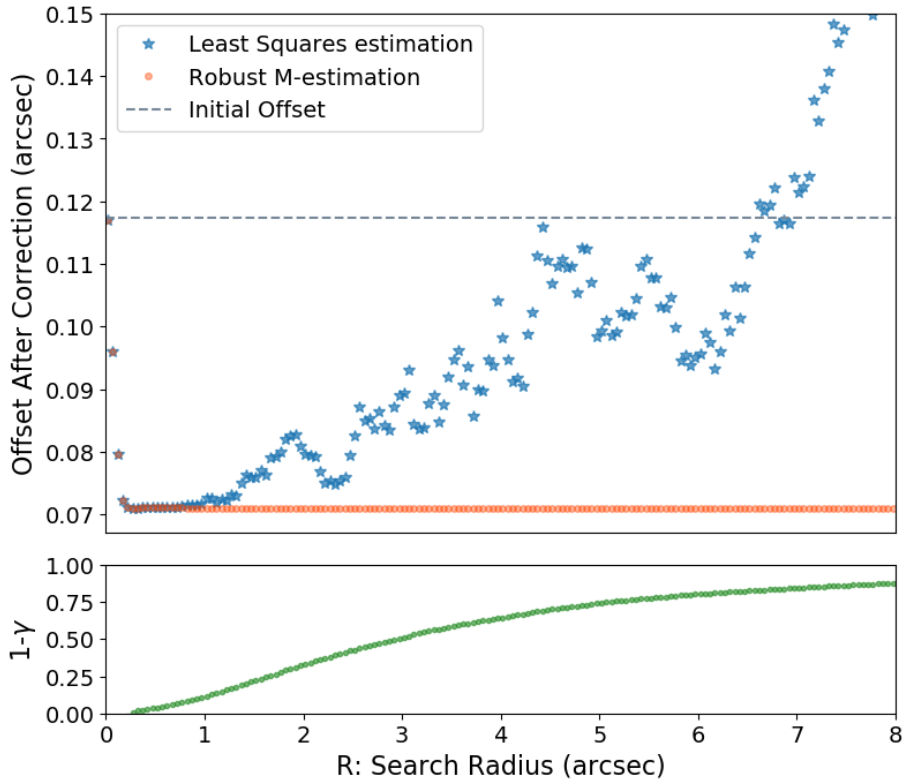

Figure 3.4: Comparison of least squares estimation with the new robust estimation tested on two images with a small offset (*grey dashed line*) and with increasing search radius. The top panel shows the offset of two images before and after correction. The bottom panel illustrates the fraction of the total number of bad matchings known from the ground truth to the total number of pairs within R.

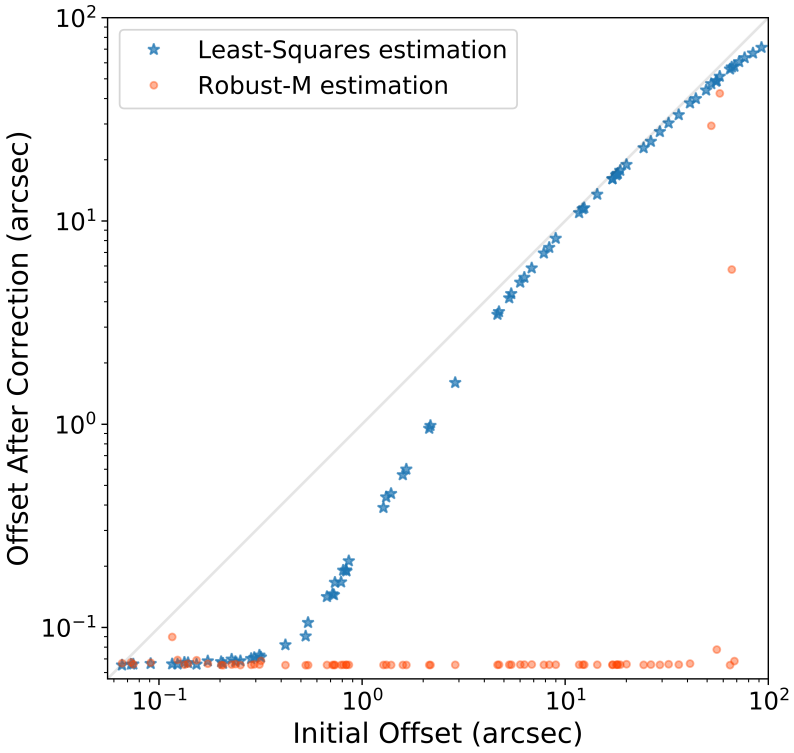

Figure 3.5: Comparison of least squares estimation with the new robust estimation on a set of simulated catalogs with increasing image offset. The diagonal line indicates the failed estimation

Chapter 4

Applications to the HST

With the success on simulations, we are now moving onto testing with the HST data. As mentioned previously, the HST is not a typical survey telescope. Instead, its observations are made based on approved programs submitted to the Space Telescope Science Institute (STScI)[1] each year and are stored in archives for public access to the data. In particular, the HLA provides on-line services to the HST data with a browsing capability of the high resolution images as well as the source lists produced from the combined images. The HSC is another archival project of the HST data. By joining and calibrating all visit-based sources lists from the HLA into a master catalog, the HSC provides a combined information on the astronomical objects detected by Hubble.

[1] See an overview from STScI at http://www.stsci.edu/hst/HST_overview

4.1 Hubble Source Catalog (HSC)

Since developed in 2016, the HSC has provided a high-quality catalog for mutipurpose use to the astronomers, but the astrometrical calibration of the HST observations is a work in progress for many aspects. One of the factors is that HST is still in orbit serving on constant program proposals. New images and data are produced each day and are added to the HLA. Other factors also include the development of the algorithms used to produce the source lists from images. Another major factor comes from the reference catalog used for calibration. In earlier versions of HSC (HSCv1 and HSCv2), the reference catalogs used to determine the absolute astrometry of the detected sources include three large catalogs: Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016), Sloan Digital Sky Servey (SDSS), and Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006). Since the absolute astrometric accuracy for these three catalogs is approximately 0.1 arcsec, the accuracy of HSC is also about 0.1 arcsec (Whitmore, Allam, et al. 2016). With the new survey telescope - the *Gaia* space telescope - launched in 2013, Version 3 of the HSC included additional calibration with *Gaia* Data Release 1 (DR1) (Gaia Collaboration et al. 2016) and provided further improved astrometry to a mode of approximately 0.003 arcsec [2]. With the increasingly available new observations and with the constant need of revisiting the calibrated sources, the

[2]See HSC Version 3 description at https://archive.stsci.edu/hst/hsc/

shortcomings of the least squares method has become in-negligible, and this is essentially the driver of developing our robust method.

4.2 Registration to *Gaia*

The *Gaia* DR2 (Lindegren et al. 2018) is released recently and has not been used for cross-calibration in HSCv3. To utilize the *Gaia* data, future versions of HSC is planning to include calibrations with the *Gaia* DR2 and here we have a chance to perform a preliminary analysis on the new data released. As a practice, our test case is a cut-out on an overlapping region of the sky covered by the HST and the *Gaia* observations. As shown in Figure (4.1), we can visualize the HSC sources and the *Gaia* DR1 and *Gaia* DR2 data as a 2D projection on the tangent plane to the celestial sphere.

As for this test case, since the systematic offset is small in this test case, our robust algorithm estimation is well performed as we can see from the result comparison plots (zoom-in view) in Figure (4.2) and Figure (4.3). In both figures, the left panel views the raw data of the two catalogs before correction and the right panel scatters the reference catalog and the corrected HST sources. As the systematic shift being recovered, the HSC catalog sources (orange dots) coincide with the majority of the *Gaia* sources (blue dots) as expected. Other sources that do not have a match are due to the different number of source

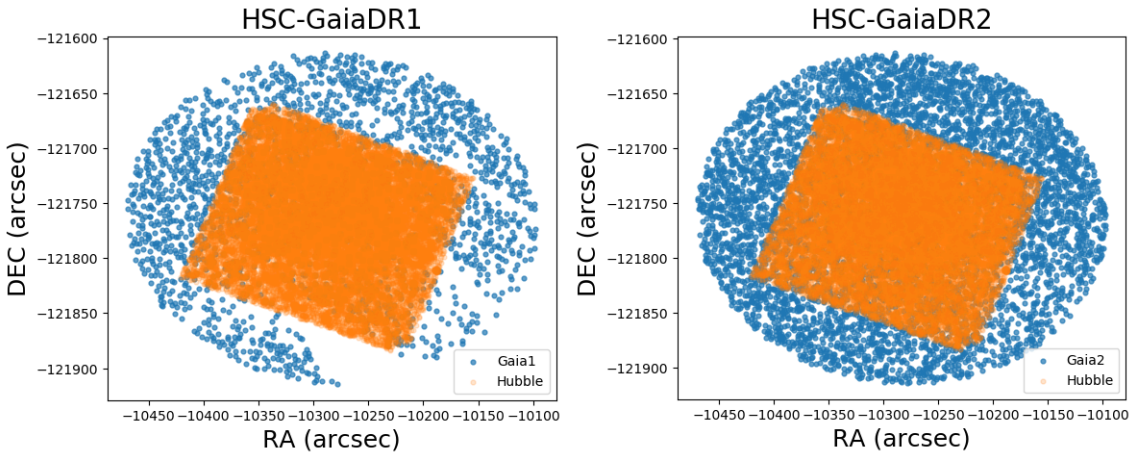

Figure 4.1: 2D Projection of HSC and *Gaia* Data

detections in two telescopes.

Figure (4.4) illustrates the pairwise residuals comparisons before and after performing the astrometric correction. As we can see from the histograms, our robust estimation has successfully recovered the shift in both pairs of catalogs. The astrometric accuracy has been improved from a mode of approximately 0.7 arcsec before correction to 0.01 arcsec after calibration.

The least squares method was also applied under the same large search radius of 5 arcsec but has failed to find the correct rotation. However, using all nearby pairs within a large search radius is not a typical procedure to work the least squares algorithm. Instead, if we have used the pre-determined set of most likely matched pairs, the least squares estimation also recovers the rotation accurately since the offset is small.

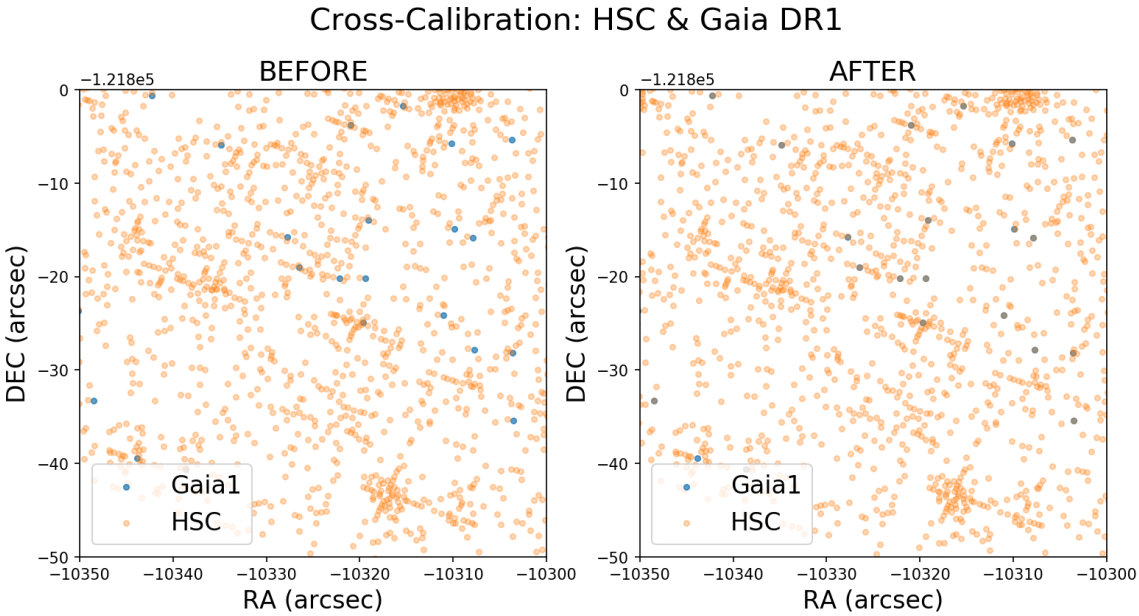

Figure 4.2: HSC to *Gaia* DR1 offset comparison before and after astrometric correction. Left panel: A little fraction of *Gaia* sources (*blue dots*) coincide with the HSC sources (*orange dots*). Right panel: The majority of the *Gaia* sources coincide with the HSC sources indicating there was a systematic shift and was successfully recovered

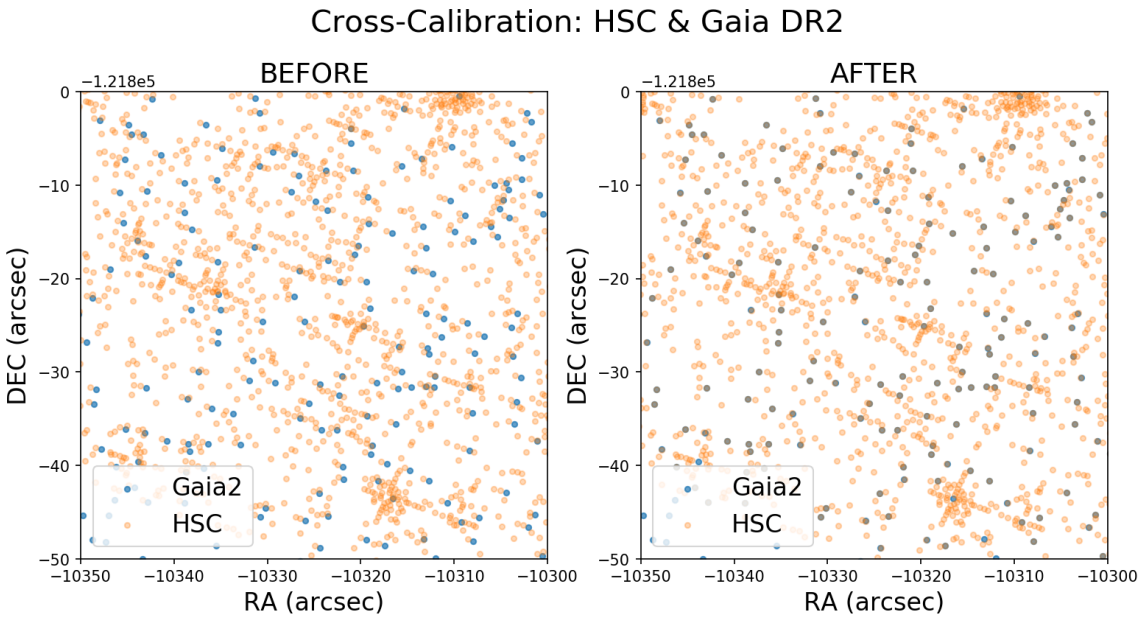

Figure 4.3: HSC to *Gaia* DR2 offset comparison before and after astrometric correction. Left panel: A systematic shift of the *Gaia* sources (*blue dots*) from the HSC sources (*orange dots*) is visible. Right panel: The shift was successfully recovered and the HSC sources are calibrated to the *Gaia* sources

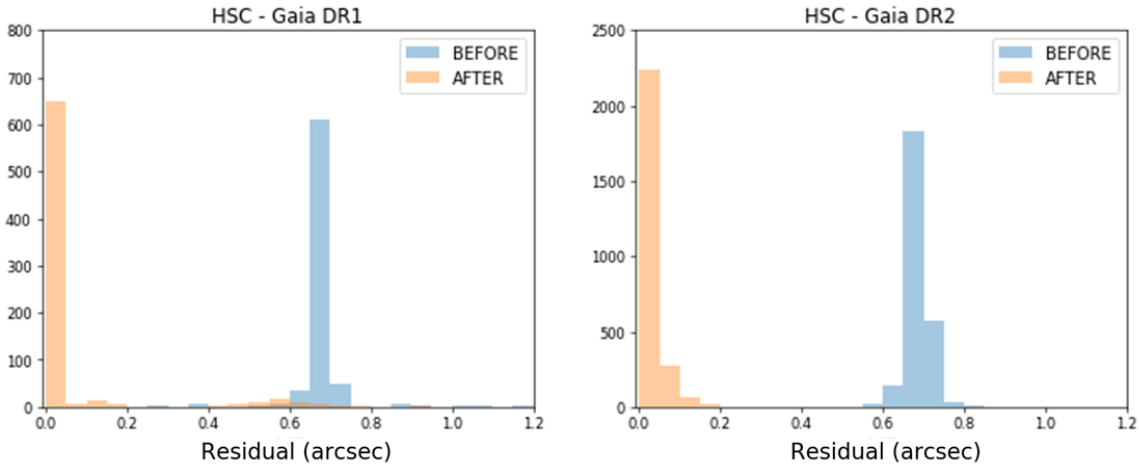

Figure 4.4: Histogram of the pairwise residuals of HSC-*Gaia*DR1 and HSC-*Gaia*DR2 before and after astrometic correction. In both cases, the accuracy has been improved from a mode of approximately 0.7 arcsec to less than a mode of 0.1 arcsec

Chapter 5

Final Remarks and Future Work

In this study, we have proposed a novel mathematical approach for cross-registering astronomical catalogs tailored to observations with a small field of view based on Bayesian framework and robust statistics. Our preliminary study on both the simulations and the real data of the HST observations have shown promising results on improving the astrometric accuracy over the state-of-art method.

However, the limitations of this algorithm should not be overlooked. In particular, the current solution to the issue of the multiple minima in the objective function associated with the σ parameter is largely intervened by artifacts. As we have formalized our objective function by approximations to be a function of residuals only and weighed down the contributions from other parameters of γ and σ, we have limited our estimation from finding a global estimate that

should have incorporated all parameters. Our future plan is therefore to revisit the likelihood function to account for effect from the other two parameters to the objective function.

Alternatively, we can always pre-determine a set of approximately matched pairs in practice by using *pre-offsets* methods described in Whitmore, Allam, et al. (2016), Bayesian likelihood comparison (Budavári and Lubow 2012), or by limiting on other information such as the magnitude of sources that are also available in the source lists. With the set of most likely associated pairs determined, our robust method is well performed as the number of bad matchings is limited.

References

Bertin, E. and S. Arnouts (1996). “SExtractor: Software for source extraction”. In: *Astronomy and Astrophysics Supplement* 117, pp. 393–404. DOI: 10.1051/aas:1996164.

Box, G. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Vol. 16. John Wiley & Sons, Ltd. ISBN: 9781118033197. DOI: 10.1002/9781118033197.

Budavári, T. and A. Basu (2016). “Probabilistic Cross-identification in Crowded Fields as an Assignment Problem”. In: *The Astronomical Journal* 152.4, p. 86. DOI: 10.3847/0004-6256/152/4/86.

Budavári, T. and T. J. Loredo (2015). “Probabilistic Record Linkage in Astronomy: Directional Cross-Identification and Beyond”. In: *Annual Review of Statistics and Its Application* 2.1, pp. 113–139. DOI: 10.1146/annurev-statistics-010814-020231.

Budavári, T. and S. H. Lubow (2012). “Catalog Matching with Astrometric Correction and its Application to the Hubble Legacy Archive”. In: *The Astrophysical Journal* 761.2, p. 188. DOI: 10.3847/0004-6256/152/4/86.

REFERENCES

Budavári, T. and A. S. Szalay (2008). "Probabilistic Cross-Identification of Astronomical Sources". In: *The Astrophysical Journal* 679.1, p. 301. DOI: 10.1086/587156.

Chambers, K. C. et al. (2016). "The Pan-STARRS1 Surveys". In: *arXiv e-prints*, arXiv:1612.05560, arXiv:1612.05560. arXiv: 1612.05560.

Fan, D. et al. (2013). "Efficient Catalog Matching with Dropout Detection". In: *Publications of the Astronomical Society of the Pacific* 125.924, p. 218. URL: http://stacks.iop.org/1538-3873/125/i=924/a=218.

Fisher, R. (1953). "Dispersion on a Sphere". In: *Proceedings of the Royal Society of London Series A* 217, pp. 295–305. DOI: 10.1098/rspa.1953.0064.

Gaia Collaboration et al. (2016). "Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties". In: *Astronomy & Astrophysics* 595, A2, A2. DOI: 10.1051/0004-6361/201629512.

Gamerman, D. et al. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC. ISBN: 9781482296426.

Greisen, E. W. and M. R. Calabretta (2002). "Representations of world coordinates in FITS". In: *Astronomy and Astrophysics* 395, pp. 1061–1075. DOI: 10.1051/0004-6361:20021326.

Heinis, S., T. Budavári, and A. S. Szalay (2009). "Cross-identification Performance from Simulated Detections: Galex and SDSS". In: *The Astrophysical Journal* 705.1, p. 739. DOI: 10.1088/0004-637X/705/1/739.

REFERENCES

Huber, P. J. (1981). *Robust Statistics*. New York, United States of America: John Wiley & Sons, Ltd. ISBN: 0471418056.

Jenkner, H. et al. (2006). "Concept for the Hubble Legacy Archive". In: *Astronomical Data Analysis Software and Systems XV*. Vol. 351. Astronomical Society of the Pacific Conference Series, p. 406. URL: http://adsabs.harvard.edu/abs/2006ASPC..351..406J.

Kerekes, G. et al. (2010). "Cross Identification of Stars with Unknown Proper Motions". In: *The Astrophysical Journal* 719.1, p. 59. DOI: 10.1088/0004-637X/719/1/59.

Lang, D. et al. (2010). "Astrometry.net: Blind Astrometric Calibration of Arbitrary Astronomical Images". In: *The Astronomical Journal* 139, pp. 1782–1800. DOI: 10.1088/0004-6256/139/5/1782.

Lindegren, L. et al. (2018). "Gaia Data Release 2. The astrometric solution". In: *Astronomy and Astrophysics* 616, A2, A2. DOI: 10.1051/0004-6361/201832727.

Lucas, R. A. et al. (2018). "ACS Data Handbook, Version 9.0". In: Baltimore: STScI. Chap. 1.

Mallinar, N., T. Budavári, and G. Lemson (2017). "Probabilistic cross-identification of galaxies with realistic clustering". In: *Astronomy and Computing* 20, pp. 83–86. DOI: 10.1016/j.ascom.2017.06.001.

REFERENCES

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust statistics theory and methods*. John Wiley & Sons, Ltd.

Marquez, M. J., T. Budavári, and L. M. Sarro (2014). "Improving cross-identification of galaxies using their photometry". In: *Astronomy and Astrophysics* 563, A14, A14. DOI: 10.1051/0004-6361/201322625.

Miller III, W., B. C. Whitmore, and H. Jenkner (2008). "Enhancing Science with a Hubble Legacy Archive". In: *Astronomical Data Analysis Software and Systems XVII*. Astronomical Society of the Pacific Conference Series, p. 478. URL: http://adsabs.harvard.edu/abs/2008ASPC..394..478M.

Shi, X., T. Budavári, and A. Basu (2017). "Probabilistic Cross-identification of Multiple Catalogs in Crowded Fields". In: *ArXiv e-prints*, arXiv:1710.10231, arXiv:1710.10231. arXiv: 1710.10231.

Skrutskie, M. F. et al. (2006). "The Two Micron All Sky Survey (2MASS)". In: *The Astronomical Journal* 131, pp. 1163–1183. DOI: 10.1086/498708.

Sorenson, H.W. (1980). *Parameter Estimation: Principles and Problems*. Control and systems theory; v.9. New York: Marcel Dekker. ISBN: 9780824769871.

Stetson, P. B. (1987). "DAOPHOT - A computer program for crowded-field stellar photometry". In: *Publications of the Astronomical Society of the Pacific* 99, pp. 191–222. DOI: 10.1086/131977.

REFERENCES

Whitmore, B. C., S. Allam, et al. (2016). "Version 1 of the Hubble Source Catalog". In: *The Astronomical Journal* 151, 134, p. 134. DOI: 10.3847/0004-6256/151/6/134.

Whitmore, B. C., K. Lindsay, and M. Stankiewicz (2008). "Source Lists for the Hubble Legacy Archive (HLA)". In: *Astronomical Data Analysis Software and Systems XVII*. Vol. 394. Astronomical Society of the Pacific Conference Series, p. 481. URL: http://adsabs.harvard.edu/abs/2008ASPC..394..481W.

York, D. G. et al. (2000). "The Sloan Digital Sky Survey: Technical summary". In: *The Astronomical Journal* 120.3, pp. 1579–1587. ISSN: 0004-6256. DOI: 10.1086/301513.

Vita

Fan Tian was born in Shaanxi, China in 1993. She received Bachelor of Environments in Civil Engineering Systems and Diploma in Mathematical Sciences from the University of Melbourne, Australia in 2016. After that, Fan joined the Johns Hopkins University for the Master of Science and Engineering degree in Applied Mathematics and Statistics from 2017 Spring. Her research includes studying cross-matching in the field of Astronomy; functional data analysis with applications to clinical studies; and theoretical analysis in multi-linear algebra.